K-Means Clustering of People with COVID-19

July 2, 2020

DoYeong, Jeon
Software Engineering Laboratory
Soongsil University

TABLE OF CONTENTS

1. Source Code

1.1. Code for Creating Database

❑ **CreatingDB Class**

```
class CreatingDB:
    """
    Class for creating random database
    """
    num_people = 0  # number of people to create
    base_date = None  # the base date of data

    def __init__(self, num_people, base_date):
        self.num_people = num_people
        self.base_date = base_date

    def generate_incurred_date(self):
        """
        function to create random incurred date
        :return:
            incurred_date: string, the day of infection or contact
            elapsed_days: int, the difference between base date and incurred
date
        """
        elapsed_days = random.randint(0, 14)  # the valid day period is 0~14
        # extracting the incurred day using periods and base date
        incurred_date = (self.base_date - timedelta(days=elapsed_days)). \
            strftime("%Y %m %d")
        return incurred_date, elapsed_days

    def generate_address_list(self):
        """
        function to get one address randomly from the adress list
        :return: the randomly generated address list
        """
        with open('./Address_Part.txt', 'r', encoding='utf-8') as add_file:
            # add_file = add_file.encoding
            address_list = add_file.readlines()

            random_address_list = []  # list to store addresses

            # extract addresses as many as the number of recipients
            for _ in range(1, self.num_people + 1):
                random_address_list.append(random.choice(address_list))

        return random_address_list

    def generate_csv_data(self):
        """
        function to create .csv file with randomly generated records
        :return: None
        """
```

```
        num_healthy = round(self.num_people / 3)  # 1/3 is healthy
        num_contacted = round(self.num_people / 3)  # 1/3 is contacted
        # 1/3 is confirmed
        num_confirmed = self.num_people - num_healthy - num_contacted

        id_list = list(range(1, self.num_people + 1))  # ID as many as people
        random.shuffle(id_list)  # shuffle list

        # age records as many as people
        age_list = list(random.randint(1, 100)
                        for _ in range(1, self.num_people + 1))
        # address records as many as people
        address_list = self.generate_address_list()

        severity_list = []  # severity records as many as people
        incurred_date_list = []  # incurred date list including 'None'(healthy)
        status_list = []  # status(Healthy, Contacted, and Confirmed) list

        # Entire people num = healthy + contacted + confirmed
        # Repeat as many healthy people
        for _ in range(num_healthy):
            # severity_list.append(0)
            status_list.append('Healthy')
            incurred_date_list.append('None')

        # Repeat as many contacted people
        for count in range(num_contacted):
            date, days = self.generate_incurred_date()
            status_list.append('Contacted')
            #     severity_list.append(round(self.compute_severity('contacted',
days), 2))
            incurred_date_list.append(date)

        # Repeat as many confirmed people
        for _ in range(num_confirmed):
            date, days = self.generate_incurred_date()
            status_list.append('Confirmed')
            #     severity_list.append(round(self.compute_severity('confirmed',
days), 2))
            incurred_date_list.append(date)

        # converting as pandas DataFrame data type to save .csv
        df = pd.DataFrame({
            "ID": id_list,
            "Age": age_list,
            "Address": address_list,
            "Covid Status": status_list,
            # "Severity": severity_list,
            "Incurred Date": incurred_date_list,
        })
        df = df.sort_values(['ID'], ascending=[True])
        df.reset_index(drop=True, inplace=True)

        # saving as .csv file
```

```
        df.to_csv("corona_data.csv", mode='w', encoding='utf-8-sig')
```

1.2. Code for Clustering

❑ **ClusteringPeople Class**

```
class ClusteringPeople:
    df_corona = None
    cluster_result_dic = {}

    def __init__(self, file_path):
        self.load_data(file_path)

    def load_data(self, file_path):
        """
        method to load .csv file
        :param file_path: string, the path of file
        :return:
        """
        self.df_corona = pd.read_csv(file_path)

    def preprocess(self):
        """
        method to preprocess the data for distance function
        :return: None
        """
        col_num = len(self.df_corona)  # the number of rows from loaded data
        today = datetime.now().date()  # date of today, YEAR-MONTH-DAY

        # selecting specific column to compute 'severity'
        incur_date_col = self.df_corona['Incurred Date']
        status = self.df_corona['Covid Status']

        severity_list = []  # list for storing severity result

        for i in range(col_num):
            severity = 0  # default is healthy, 0.
            if status[i] == 'Contacted':  # contacted person?
                # formula for contacted person:
                #   x = 1 - ((today's date) - (infected date)) * 0.05)
                elapsed_days = (today - parse(incur_date_col[i]).date()).days
                severity = 1 - (elapsed_days * 0.05)
            elif status[i] == 'Confirmed':  # confirmed person?
                # formula for confirmed person:
                #   x = (1 - ((today's date) - (infected date)) * 0.05)) / 2
                elapsed_days = (today - parse(incur_date_col[i]).date()).days
                severity = (1 - (elapsed_days * 0.05)) * 0.5

            severity_list.append(severity)  # add the value to the list
        self.df_corona["Severity"] = severity_list

    def cluster(self):
```

```
        sse_list = []  # list for storing SSE(Sum of squares errors)
        silhouette_score_list = []  # list for storing silhouette scores

        for i in range(2, 10):  # number of clusters 2 to 9
            # load the k-means model
            km = cluster.KMeans(n_clusters=i,  # the number of cluster
                                init='k-means++',    # how to initial cluster
centers
                                max_iter=300,  # maximum number of iterations
                                algorithm='auto'   # three choices: auto, full,
and elkan.
                                )

            # changing the shape of data
            severity_list = self.df_corona["Severity"].values.tolist()
            severity_list = np.array(severity_list)

            # cluster
            cluster_predicted_list  =  km.fit_predict(severity_list.reshape(-1,
1))

            # storing SSE value to get the optimal number of cluster
            sse_list.append(km.inertia_)

            # storing silhouette score to get optimal number of cluster

silhouette_score_list.append(silhouette_score(severity_list.reshape(-1,    1),
cluster_predicted_list))

            cluster_list = [j for j in range(i)]  # cluster list
            # display the reuslt of cluster
            self.print_result_of_cluster(cluster_list, cluster_predicted_list)

            # store the prediction result
            self.cluster_result_dic[i] = cluster_predicted_list

    def draw_elbow_method(self, sse_list):
        """
        method to draw elbow graph using SSE(Sum of Squares Error)
        :param sse_list: list of SSE
        :return: None
        """
        plt.plot(range(2, 10), sse_list, marker='o')
        plt.xlabel("The Number of Cluster")
        plt.ylabel("SSE")
        plt.show()

    def print_result_of_cluster(self, cluster_list, cluster_predicted_list):
        """
        form
        Cluster 1:
            Number of people: n
            Severity Values: [...]
            Average of severities: n
```

```
        Cluster 2:
            ...

        :return:
        """
        severity_list = self.df_corona["Severity"].values.tolist()
        cluster_predicted_list = cluster_predicted_list.tolist()
        print(f"The number of Cluster: {len(cluster_list)}")
        for cluster_idx in cluster_list:
            num_people = cluster_predicted_list.count(cluster_idx)
            cluster_severity_list = []
            for person_idx in range(len(cluster_predicted_list)):
                if cluster_idx == cluster_predicted_list[person_idx]:

cluster_severity_list.append(round(severity_list[person_idx], 2))
            print(f"\tCluster {cluster_idx}:")
            print(f"\t\tNumber of Poeple: {num_people}")
            print(f"\t\tSeverity Values: {cluster_severity_list}")
            print(f"\t\tAverage                    of                    severities:
{round(sum(cluster_severity_list) / len(cluster_severity_list), 2)}")
        print()  # float 1 line


    def draw_silhouette(self):
        """
        method to draw graph using silhouette scores
        :return: None
        """
        pass

    def draw_graph(self):
        """
        method to draw clustering result
        :return: None
        """
        pass
```

- **main**

```
if __name__ == '__main__':
    # CODE FOR CLUSTERING
    file_path = './corona_data.csv'

    cp = ClusteringPeople(file_path)
    cp.preprocess()
    cp.draw_graph()
    cp.cluster()
```

2. Result of Clustering

2.1. K-Means

- **The number of Cluster: 2**

```
The number of Cluster: 2
    Cluster 0:
        Number of Poeple: 57
        Severity Values: [0.3, 0.0, 0.3, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.28, 0.0, 0.25, 0.0,
  0.0, 0.0, 0.0, 0.25, 0.22, 0.17, 0.0, 0.17, 0.0, 0.0, 0.3, 0.0, 0.32, 0.0, 0.0, 0.0, 0.0, 0.17, 0.0,
  0.3, 0.25, 0.28, 0.25, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.3, 0.32, 0.0, 0.3, 0.15, 0.15, 0.0, 0.0, 0.0,
  0.0, 0.15, 0.0, 0.3]
        Average of severities: 0.1
    Cluster 1:
        Number of Poeple: 43
        Severity Values: [0.45, 0.5, 0.7, 0.4, 0.9, 0.35, 0.65, 0.47, 0.6, 0.47, 0.6, 0.8, 0.4, 0.47,
 0.65, 0.5, 0.95, 0.4, 0.4, 0.8, 0.47, 0.5, 0.95, 0.65, 0.45, 0.6, 0.38, 0.6, 0.45, 0.6, 0.5, 0.8,
 0.7, 0.7, 0.8, 0.95, 0.45, 0.45, 0.35, 0.35, 0.7, 0.35, 0.35]
        Average of severities: 0.57
```

❑ The number of Cluster: 3

```
The number of Cluster: 3
    Cluster 0:
        Number of Poeple: 41
        Severity Values: [0.3, 0.45, 0.5, 0.3, 0.2, 0.4, 0.35, 0.47, 0.28, 0.25, 0.47, 0.25, 0.22,
 0.4, 0.47, 0.5, 0.3, 0.4, 0.32, 0.4, 0.47, 0.5, 0.45, 0.3, 0.25, 0.38, 0.45, 0.28, 0.25, 0.5, 0.2,
 0.3, 0.45, 0.45, 0.32, 0.35, 0.3, 0.35, 0.35, 0.3, 0.35]
        Average of severities: 0.36
    Cluster 1:
        Number of Poeple: 20
        Severity Values: [0.7, 0.9, 0.65, 0.6, 0.6, 0.8, 0.65, 0.95, 0.8, 0.95, 0.65, 0.6, 0.6, 0.6,
 0.8, 0.7, 0.7, 0.8, 0.95, 0.7]
        Average of severities: 0.73
    Cluster 2:
        Number of Poeple: 39
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.17, 0.0,
 0.17, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.17, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.15, 0.15, 0.0,
 0.0, 0.0, 0.0, 0.15, 0.0]
        Average of severities: 0.02
```

❑ The number of Cluster: 4

```
The number of Cluster: 4
    Cluster 0:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 1:
        Number of Poeple: 22
        Severity Values: [0.45, 0.5, 0.4, 0.47, 0.6, 0.47, 0.6, 0.4, 0.47, 0.5, 0.4, 0.4, 0.47, 0.5,
 0.45, 0.6, 0.6, 0.45, 0.6, 0.5, 0.45, 0.45]
        Average of severities: 0.49
    Cluster 2:
        Number of Poeple: 15
        Severity Values: [0.7, 0.9, 0.65, 0.8, 0.65, 0.95, 0.8, 0.95, 0.65, 0.8, 0.7, 0.7, 0.8, 0.95,
 0.7]
        Average of severities: 0.78
    Cluster 3:
        Number of Poeple: 30
        Severity Values: [0.3, 0.3, 0.2, 0.35, 0.28, 0.25, 0.25, 0.22, 0.17, 0.17, 0.3, 0.32, 0.17,
 0.3, 0.25, 0.38, 0.28, 0.25, 0.2, 0.3, 0.32, 0.35, 0.3, 0.15, 0.15, 0.35, 0.15, 0.35, 0.3, 0.35]
        Average of severities: 0.27
```

❑ The number of Cluster: 5

```
The number of Cluster: 5
    Cluster 0:
        Number of Poeple: 24
        Severity Values: [0.3, 0.3, 0.2, 0.28, 0.25, 0.25, 0.22, 0.17, 0.17, 0.3, 0.32, 0.17, 0.3,
 0.25, 0.28, 0.25, 0.2, 0.3, 0.32, 0.3, 0.15, 0.15, 0.15, 0.3]
        Average of severities: 0.25
    Cluster 1:
        Number of Poeple: 12
        Severity Values: [0.7, 0.65, 0.6, 0.6, 0.65, 0.65, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7]
        Average of severities: 0.65
    Cluster 2:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 3:
        Number of Poeple: 8
        Severity Values: [0.9, 0.8, 0.95, 0.8, 0.95, 0.8, 0.8, 0.95]
        Average of severities: 0.87
    Cluster 4:
        Number of Poeple: 23
        Severity Values: [0.45, 0.5, 0.4, 0.35, 0.47, 0.47, 0.4, 0.47, 0.5, 0.4, 0.4, 0.47, 0.5, 0.45,
  0.38, 0.45, 0.5, 0.45, 0.45, 0.35, 0.35, 0.35, 0.35]
        Average of severities: 0.43
```

- **The number of Cluster: 6**

```
The number of Cluster: 6
    Cluster 0:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 1:
        Number of Poeple: 12
        Severity Values: [0.7, 0.65, 0.6, 0.6, 0.65, 0.65, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7]
        Average of severities: 0.65
    Cluster 2:
        Number of Poeple: 21
        Severity Values: [0.3, 0.3, 0.35, 0.28, 0.25, 0.25, 0.3, 0.32, 0.3, 0.25, 0.38, 0.28, 0.25,
 0.3, 0.32, 0.35, 0.3, 0.35, 0.35, 0.3, 0.35]
        Average of severities: 0.31
    Cluster 3:
        Number of Poeple: 8
        Severity Values: [0.9, 0.8, 0.95, 0.8, 0.95, 0.8, 0.8, 0.95]
        Average of severities: 0.87
    Cluster 4:
        Number of Poeple: 17
        Severity Values: [0.45, 0.5, 0.4, 0.47, 0.47, 0.4, 0.47, 0.5, 0.4, 0.4, 0.47, 0.5, 0.45, 0.45,
  0.5, 0.45, 0.45]
        Average of severities: 0.45
    Cluster 5:
        Number of Poeple: 9
        Severity Values: [0.2, 0.22, 0.17, 0.17, 0.17, 0.2, 0.15, 0.15, 0.15]
        Average of severities: 0.18
```

❑ The number of Cluster: 7

```
The number of Cluster: 7
    Cluster 0:
        Number of Poeple: 13
        Severity Values: [0.2, 0.25, 0.25, 0.22, 0.17, 0.17, 0.17, 0.25, 0.25, 0.2, 0.15, 0.15, 0.15]
        Average of severities: 0.2
    Cluster 1:
        Number of Poeple: 12
        Severity Values: [0.7, 0.65, 0.6, 0.6, 0.65, 0.65, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7]
        Average of severities: 0.65
    Cluster 2:
        Number of Poeple: 17
        Severity Values: [0.45, 0.5, 0.4, 0.47, 0.47, 0.4, 0.47, 0.5, 0.4, 0.4, 0.47, 0.5, 0.45, 0.45,
  0.5, 0.45, 0.45]
        Average of severities: 0.45
    Cluster 3:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 4:
        Number of Poeple: 4
        Severity Values: [0.8, 0.8, 0.8, 0.8]
        Average of severities: 0.8
    Cluster 5:
        Number of Poeple: 17
        Severity Values: [0.3, 0.3, 0.35, 0.28, 0.3, 0.32, 0.3, 0.38, 0.28, 0.3, 0.32, 0.35, 0.3,
 0.35, 0.35, 0.3, 0.35]
        Average of severities: 0.32
    Cluster 6:
        Number of Poeple: 4
        Severity Values: [0.9, 0.95, 0.95, 0.95]
        Average of severities: 0.94
```

❑ The number of Cluster: 8

```
The number of Cluster: 8
    Cluster 0:
        Number of Poeple: 13
        Severity Values: [0.45, 0.5, 0.47, 0.47, 0.47, 0.5, 0.47, 0.5, 0.45, 0.45, 0.5, 0.45, 0.45]
        Average of severities: 0.47
    Cluster 1:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
  0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 2:
        Number of Poeple: 4
        Severity Values: [0.8, 0.8, 0.8, 0.8]
        Average of severities: 0.8
    Cluster 3:
        Number of Poeple: 14
        Severity Values: [0.3, 0.3, 0.28, 0.25, 0.25, 0.22, 0.3, 0.3, 0.25, 0.28, 0.25, 0.3, 0.3, 0.3]
        Average of severities: 0.28
    Cluster 4:
        Number of Poeple: 12
        Severity Values: [0.7, 0.65, 0.6, 0.6, 0.65, 0.65, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7]
        Average of severities: 0.65
    Cluster 5:
        Number of Poeple: 12
        Severity Values: [0.4, 0.35, 0.4, 0.4, 0.32, 0.4, 0.38, 0.32, 0.35, 0.35, 0.35, 0.35]
        Average of severities: 0.36
    Cluster 6:
        Number of Poeple: 8
        Severity Values: [0.2, 0.17, 0.17, 0.17, 0.2, 0.15, 0.15, 0.15]
        Average of severities: 0.17
    Cluster 7:
        Number of Poeple: 4
        Severity Values: [0.9, 0.95, 0.95, 0.95]
        Average of severities: 0.94
```

- **The number of Cluster: 9**

```
The number of Cluster: 9
    Cluster 0:
        Number of Poeple: 33
        Severity Values: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        Average of severities: 0.0
    Cluster 1:
        Number of Poeple: 8
        Severity Values: [0.65, 0.6, 0.6, 0.65, 0.65, 0.6, 0.6, 0.6]
        Average of severities: 0.62
    Cluster 2:
        Number of Poeple: 15
        Severity Values: [0.3, 0.3, 0.28, 0.25, 0.25, 0.3, 0.32, 0.3, 0.25, 0.28, 0.25, 0.3, 0.32,
 0.3, 0.3]
        Average of severities: 0.29
    Cluster 3:
        Number of Poeple: 4
        Severity Values: [0.9, 0.95, 0.95, 0.95]
        Average of severities: 0.94
    Cluster 4:
        Number of Poeple: 13
        Severity Values: [0.45, 0.5, 0.47, 0.47, 0.47, 0.5, 0.47, 0.5, 0.45, 0.45, 0.5, 0.45, 0.45]
        Average of severities: 0.47
    Cluster 5:
        Number of Poeple: 9
        Severity Values: [0.2, 0.22, 0.17, 0.17, 0.17, 0.2, 0.15, 0.15, 0.15]
        Average of severities: 0.18
    Cluster 6:
        Number of Poeple: 10
        Severity Values: [0.4, 0.35, 0.4, 0.4, 0.4, 0.38, 0.35, 0.35, 0.35, 0.35]
        Average of severities: 0.37
    Cluster 7:
        Number of Poeple: 4
        Severity Values: [0.8, 0.8, 0.8, 0.8]
        Average of severities: 0.8
    Cluster 8:
        Number of Poeple: 4
        Severity Values: [0.7, 0.7, 0.7, 0.7]
        Average of severities: 0.7
```